Test the Overall Significance of p-values by Using Joint Tail Probability of Ordered p-values as Test Statistic

Fang, Yongxiang; Wit, Ernst

Published in:
Advanced Data Mining and Applications

Publication date:
2008

Test the Overall Significance of p-values by Using Joint Tail Probability of Ordered p-values as Test Statistic

Yongxiang Fang* and Ernst Wit

Department of Mathematics and statistics
Lancaster University
Lancaster, LA1 4YF, UK
y.fang@lancaster.ac.uk

Abstract. Fisher's combined probability test is the most commonly used method to test the overall significance of a set independent p-values. However, it is very obviously that Fisher's statistic is more sensitive to smaller p-values than to larger p-value and a small p-value may overrule the other p-values and decide the test result. This is, in some cases, viewed as a flaw. In order to overcome this flaw and improve the power of the test, the joint tail probability of a set p-values is proposed as a new statistic to combine and make an overall test of the p-values. Through the development of a method and a practical application, this study reveals that the new method has plausible properties and more power.

Keywords: p-values, joint tail probability, combined probability test.

1 Introduction

The p-value is the probability of obtaining a value of the test statistic at least as extreme as the one that was actually observed, given that the null hypothesis is true. In many cases of statistical analysis, taking the p-values from a set of independent hypothesis tests as statistics and combining their significance is required. There are several methods for combining p-values available, for example Fisher's combined probability test [1], minimum p-value test [2], sum p-value method [3], [4], Wilkinson method [5], and inverse normal method [6]. There are also some review and comparative studies of methods for combining p-values. For example, Birnbaum [7], Littell and Folks [8], [9] and Berk and Cohen [10]. These studies essentially agree that Fisher's method is generally best and efficient among the methods mentioned above, however, none of the methods are uniformly more powerful then the others and different methods are usually sensitive in different pattern of outliers.

* Please note that the LNCS Editorial assumes that all authors have used the western naming convention, with given names preceding surnames. This determines the structure of the names in the running heads and the author index.

C. Tang et al. (Eds.): ADMA 2008, LNAI 5139, pp. 435–443, 2008.

Fisher's statistic is $T_{fk} = -2ln(\Pi_{i=1}^{k} p_i)$, while k is the number of p-values. Since the statistic is a logarithm transformed product of p-values, it is more sensitive to small p-value than large p-value. This property is viewed as a flaw especially in certain applications [11]. Furthermore, Fisher's statistic is actually also a transformed joint tail probability of the p-values, because the p-values are supposed from independent hypothesis tests. Therefore, we introduce and discuss a slightly different statistic from Fisher's for combining p-values, The new statistic is defined as the joint tail probability of **ordered p-values**. The value of Fisher's statistic depends only on the product or geometric mean of a set of p-values, no matter how alike or different between the individual elements. Our statistic is different from this. To say there are two different sets of p-values with the same product, in the first set of p-values all the elements have the same value, in the second set of p-values the elements have different values, then the the first set of p-values will be valued smaller than the second set of p-values under our statistic. Therefore, we expect that testing on our statistic shows some plausible properties and higher power.

2 Methods

Let's start from the computation of our statistic for given p-values. Denote by $p_1, p_2, ..., p_k$ k p-values from independent hypothesis tests. Let $p_{[1]} \leq p_{[2]} \leq ... \leq p_{[k]}$ be the ordered k p-values and V_k the joint tail probability of the ordered k p-values, V_k can be expressed by:

$$V_k = k! \int_0^{p_{[1]}} \int_{x_1}^{p_{[2]}} ... \int_{x_{k-1}}^{p_{[k]}} dx_k dx_{k-1} ... dx_1 \tag{1}$$

Let $V_0 = 1$ and V_j (j=1, 2,...,k) be the joint tail probability of $p_{[k-j+1]}, p_{[k-j+2]}$, $..., p_{[k]}$, obviously $V_1 = p_{[k]}$. We can have:

$$V_j = \sum_{i=1}^{j} (-1)^{i-1} \binom{j}{i} p_{[k-j+1]}^{i} V_{j-i} \tag{2}$$

Proof. When $j = 1$ equation (2) is correct, because it simply becomes $V_1 = p_{[k]}$. Suppose when $j = n < k$ equation (2) stands. For $j = n + 1$, we have

$$V_{n+1} = (n+1)! \int_0^{p_{[k-n]}} \int_{x_{k-n}}^{p_{[k-n+1]}} ... \int_{x_{k-1}}^{p_{[k]}} dx_k dx_{k-1} ... dx_{k-n}$$

and it can be written as:

$$\begin{aligned} V_{n+1} = (n+1) \int_0^{p_{[k-n]}} (n! \int_0^{p_{[k-n+1]}} ... \int_{x_{k-1}}^{p_{[k]}} dx_k dx_{k-1} ... d_{k-n+1}) dx_{k-n} \\ -(n+1) \int_0^{p_{[k-n]}} (n! \int_0^{x_{k-n}} ... \int_{x_{k-1}}^{p_{[k]}} dx_k dx_{k-1} ... d_{k-n+1}) dx_{k-n} \end{aligned}$$

Making use of the definition of V_j and equation (2), we can have:

$$V_{n+1} = (n+1)p_{[k-n]}V_n - (n+1)\int_0^{p_{[k-n]}} (\sum_{i=1}^{n} (-1)^{i-1} \binom{n}{i} x_{k-n}^i V_{n-i}) dx_{k-n}$$

On integration and simplification we have

$$V_{n+1} = \sum_{i=1}^{n+1} (-1)^{i-1} \binom{n+1}{i} p_{[k-(n+1)+1]}^i V_{(n+1)-i}$$

Equation (2) is proved.

Hence for a given set of p-values, the value of our statistic can be computed recursively by using equation (2). The computation starts with *n=1* and increases n by 1 for a new iteration until *n=k*. V_k as a statistic has its own probability density function (p.d.f) and cumulative probability function (c.d.f). There are two useful properties about V_k and its c.d.f $F(x) = Pr(V_k \leq x \mid null)$. First one is obvious $0 \leq V_k \leq 1$, because V_k is defined as joint tail probability of ordered k p-values. The second is $F(x) \geq x$, because the equation below:

$$k! \int_0^{1-\sqrt[k]{1-x}} \int_{x_1}^{1} ... \int_{x_k}^{1} dx_k dx_{k-1} ... dx_1 = x \tag{3}$$

In fact, let $\mho$ be the collection of all the ordered p-value sets which satisfy $V_k \leq x$, then integration limits of the definite integral in the left hand side of equation (3) pick out a subset $\complement$ from the collection $\mho$. That is under null hypotheses, the chance for a set of ordered p-values P_o being a member of $\complement$ is $Pr(V_k \leq x | P_o \in \complement) = x$. Therefore, $F(x) \geq x$ stands.

Now we discuss the distribution function for our statistic. The simplest case is when there are two p-values. Based on equation (2) and by some simplification, the statistic can be formulated as:

$$V_2 = 2p_{[1]}p_{[2]} - p_{[1]}^2 \tag{4}$$

Based on equation (4), we can transfer $V_2 \leq x$ for $(0 \leq x \leq 1)$ into constraints to $p_{[1]}$ and $p_{[2]}$. The constraint of $p_{[1]}$ is obviously that:

$$0 \leq p_{[1]} \leq \sqrt{x} \tag{5}$$

There are two different constants on $p_{[2]}$ when equation (5) is satisfied.

$$p_{[1]} < p_{[2]} \leq 1 \ \ when \ \ 0 < p_{[1]} \leq 1 - \sqrt{1-x} \tag{6}$$

$$p_{[1]} < p_{[2]} \leq \frac{x + p_{[1]}^2}{2p_{[1]}} \ \ when \ \ 1 - \sqrt{1-x} < p_{[1]} \leq \sqrt{x} \tag{7}$$

Hence the cumulative distribution function of our statistic for *k=2* can be worked out through integrals below:

$$Pr(V_2 \leq x) = 2\int_0^{1-\sqrt{1-x}} \int_{x_1}^{1} dx_2 dx_1 + 2\int_{1-\sqrt{1-x}}^{\sqrt{x}} \int_{x_1}^{\frac{x+x_1^2}{2x_1}} dx_2 dx_1 \quad (8)$$

From equation (8) we have the c.d.f of our statistic for k=2.

$$F(x) = 1 - \sqrt{1-x} + xln\frac{\sqrt{x}}{1-\sqrt{1-x}} \quad (9)$$

Hence we have its corresponding p.d.f:

$$f(x) = ln\frac{\sqrt{x}}{1-\sqrt{1-x}} \quad (10)$$

Due to V_2 is the tail probability of given two ordered p-values and equation (9) is its distribution function, replacing x in equation (9) with V_2 produces actually the extremity of a set of two ordered p-values. That is to put the value of V_2 into equation (9), we directly get the p-value of testing on V_2.

For $k = 3$, we have

$$V_3 = 6p_{[1]}p_{[2]}p_{[3]} - 3p_{[1]}p_{[2]}^2 - 3p_{[1]}^2p_{[3]} + p_{[1]}^3 \quad (11)$$

Similarly, from equation (11) we can transfer $V_3 \leq x$ into constraints to $p_{[1]}$, $p_{[2]}$ and $p_{[3]}$ and formulate the probability distribution V_3 as the sum of four integrals:

$$F(x) = Pr(V_3 < x) = I_1 + I_2 + I_3 + I_4 \quad (12)$$

while $I_1 = 6\int_0^{b_{11}} \int_{x_1}^1 \int_{x_2}^1 dx_3 dx_2 dx_1$, $I_2 = 6\int_{b_{11}}^{b_{12}} \int_{x_1}^{b_{21}} \int_{x_2}^1 dx_3 dx_2 dx_1$, $I_3 = 6\int_{b_{11}}^{b_{12}} \int_{b_{21}}^{b_{22}} \int_{x_2}^{b_{31}} dx_3 dx_2 dx_1$, $I_4 = 6\int_{b_{12}}^{b_{13}} \int_{x_1}^{b_{22}} \int_{x_2}^{b_{31}} dx_3 dx_2 dx_1$.

while $b_{11} = 1 - \sqrt[3]{1-x}$, b_{12} **is solution of** $b_{12}^3 - \frac{3}{2}b_{12}^2 + \frac{x}{2} = 0$ **in [0,1]**, $b_{13} = \sqrt[3]{x}$, $b_{21} = 1 - \sqrt{1 + \frac{x_1^3 - 3x_1^2 - x}{3x_1}}$, $b_{22} = \frac{1}{2}(1 + \sqrt{\frac{4x - x_1^3}{3x_1}})$, $b_{31} = \frac{x + 3x_1x_2^3 - x_1^3}{6x_1x_2 - 3x_1^2}$.

Due to there is no analytical solution for the integral I_3, the computation of $F(V_3)$ (the c.d.f of V_3) cannot be done without numeric integration. This makes the test on V_3 complicate. Obviously, when the number of p-values $k > 3$, the test on our statistic becomes more difficult. Therefore, how to test on our statistic in a simple way is the problem should be first solved in the practical application of our statistic.

A solution can be developed based on two properties of V_k which are $0 \leq V_k \leq 1$ and $Pr(V_k \leq x \mid null) \geq x$. Denote by α the significance level and $V_{k,\alpha}$ the critical value, then a power scale $\gamma_{k,\alpha}$ for V_k exists so that:

$$Pr(V_k \leq V_{k,\alpha} \mid null) = Pr(V_k^{\gamma_{k,\alpha}} \leq \alpha \mid null) = \alpha \quad (13)$$

Table 1. $\gamma_{k,\alpha}$values

k	$\gamma_{k,0.01}$	$\gamma_{k,0.05}$	k	$\gamma_{k,0.01}$	$\gamma_{k,0.05}$
1	1	1	16	0.4069	0.3452
2	0.762	0.7159	17	0.4016	0.3400
3	0.6602	0.5995	18	0.3970	0.3350
4	0.5982	0.5368	19	0.3929	0.3306
5	0.5560	0.4935	20	0.3890	0.3264
6	0.5244	0.4635	21	0.3852	0.3224
7	0.5023	0.4395	22	0.3816	0.3187
8	0.4848	0.4204	23	0.3783	0.3152
9	0.4697	0.4062	24	0.3749	0.3121
10	0.4569	0.3937	25	0.3715	0.3092
11	0.4460	0.3830	26	0.3682	0.3064
12	0.4365	0.3735	27	0.3647	0.3038
13	0.4276	0.3653	28	0.3615	0.3014
14	0.4201	0.358	29	0.3587	0.2992
15	0.4131	0.3514	30	0.3561	0.2971

Table 2. Simulated rejecting rate to null cases

k	$P^*_{0.01}$	CI of $P^*_{0.01}$	$P^*_{0.05}$	CI of $P^*_{0.05}$
1	0.0117	(0.0095799, 0.013820)	0.0504	(0.048280, 0.052520)
2	0.0112	(0.0091257, 0.013274)	0.0517	(0.049626, 0.053774)
3	0.0086	(0.0067824, 0.010418)	0.0490	(0.047182, 0.050818)
4	0.0100	(0.0080400, 0.011960)	0.0516	(0.049640, 0.053560)
5	0.0110	(0.0089443, 0.013056)	0.0490	(0.046944, 0.051056)
6	0.0098	(0.0078597, 0.011740)	0.0449	(0.042960, 0.046840)
7	0.0097	(0.0077696, 0.011630)	0.0499	(0.047970, 0.051830)
8	0.0106	(0.0085821, 0.012618)	0.0508	(0.048782, 0.052818)
9	0.0094	(0.0074997, 0.011300)	0.0493	(0.047400, 0.051200)
10	0.0101	(0.0081302, 0.012070)	0.0513	(0.046861, 0.055739)

From equation (14) we have:

$$\gamma_{k,\alpha} = \frac{\ln(\alpha)}{\ln(V_{k,\alpha})} \tag{14}$$

Based on equation (13) and (14), we can simplify the test as a simple comparing the power scaled our statistic with the significance level, if the critical value $V_{k,\alpha}$ is available. Due to the c.d.f of V_k (when $k \geq 3$) is not analytical, we proposed to work out $V_{k,\alpha}$ using simulation technique and then get $\gamma_{k\alpha}$ from equation (14). The simulation is simply drawing a large number N sets of p-values from $U(0,1)^k$ and computing V_k's value from each set of them. Order these values and take the one being ordered as $N \times \alpha$ to be a realization of $V_{k,\alpha}$. The process is repeated M times then $\hat{V}_{k,\alpha}$ is valued by the mean of M realizations.

Our simulation takes N=10000 and M=100, for k=3 to k=30, $\alpha = 0.05$ and $\alpha = 0.01$, the results are shown in table 1. Where for *k=1* and *k=2* the power

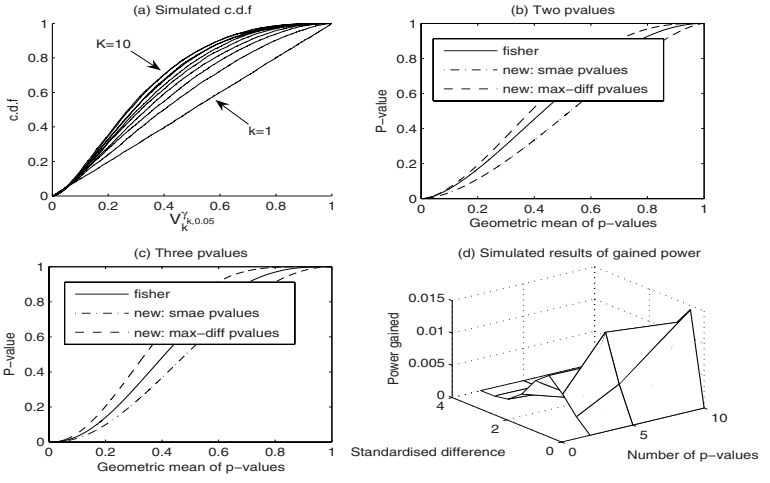

Fig. 1. Results from methods comparisons

scale is computed from the corresponding distribution function, therefore no confidence interval is required.

A simulation of the application of the proposed technique is conducted for k = 1,2, ..., 10 and $\alpha = 0.05$. For each value of k, 10000 sets of p-values are drawn from $U(0,1)^k$, then V_k is computed and it is power scaled by corresponding $\gamma_{k,0.01}$ (and $\gamma_{k,0.05}$) so that to get $V_k^{\gamma_{k,\alpha}}$. The $V_k^{\gamma_{k,\alpha}}$ with values smaller than α are the null hypotheses should be rejected. For the cases in this simulation, the expected rejecting rate is α. The simulated rates of rejection are shown in Table 2, which are very encouraging on comparing them with the corresponding α. We plot the simulated c.d.f of the power scaled our statistic in Figure 1 subplot(a) for $\alpha = 0.05$. It shows that for all $k = 1, 2, ..., 10$, when $x < 0.05$, $Pr(V_k^{\gamma_{k,\alpha}} \leq x) < x$ and $Pr(V_k^{\gamma_{k,\alpha}} \leq x) \simeq x$. when $x > 0.05$, $Pr(V_k^{\gamma_{k,\alpha}} \leq x) > x$.

3 Discussion

We expected our method has some plausible behaviors which pictures the difference from Fisher's combined probability test, hence a comparison is made in this section. Let $\wp$ contain all different sets of p-values whose products equal a given value $c \in (0,1)$. Let $V_k(\wp)$ be collection of their corresponding values under our statistic. Hance, a set of p-values whose all elements have the same value $\sqrt[k]{c}$ will be in $\wp$ and its corresponding value under our statistic will be the smallest one in $V_k(\wp)$. On the other hand, if a set of p-values has the biggest variation among their elements and belongs $\wp$, combining this set of p-values under our approach, will produce the biggest value in $V_k(\wp)$.

When k=2, the two sets of p-values being the smallest and largest under our statistic are correspondingly $(\sqrt{c},\sqrt{c})$ and $(1-\sqrt{(1-c)},1)$. We test them by

Fisher's method and ours. By changing c's value we get three sets results, one is from Fisher's, the other two are from our method. They are illustrated in Figure 1 subplot (b). Similar work is done for number of p-values k=3, and the results is shown in Figure 1 subplot (c). From the two subplots we can see that if two sets of p-values are with the same value of product they are tested no difference under Fisher's test. However, the one which has smaller variation among their elements will be tested with smaller p-values under our test. In another words, our method is more sensitive to p-value sets which their elements are uniformly small. This means our test are more plausible for testing where the null hypotheses are commonly true in the k independent tests.

A simulation based comparison between Fisher's with ours confirms the above analysis. In the simulation, we suppose the variation is normally distributed, the null hypothesis is that the original individual test statistic is with zero mean, the alternative hypothesis is that the mean of the original test statistics moves away from zero and the distance of the move is scaled by standard deviation. We select the distance as $0.5 \times d$ where $d = 1, 2, ..., 8$. The number of p-values are 2, 5 and 10 respectively and the significance level α are 0.01 and 0.05 respectively. The results shows our method is more powerful than Fisher's. We present the power gained by our method, compared to Fisher's, in Figure 1 subplot(d).

Finally, we make two comparisons by using practical examples. The first is a simple one which has been used by William Rice [11]. A biological study were conducted twice in two different years respectively, and two p-values are 1/120 and 1 respectively. The product of he two p-values is 1/120 and using Fisher's method produces a p-values 0.049 which is significant at 0.05 level. Due to the two element are extremely different, under our method the test results shows it is not significant at 0.05. level, because the p-value from our test is 0.0538.

The second example is an microarray study on the effects of treating roach with chemicals of three different concentrations 0.1 ng EE2/L, 1 ng EE2/L and 10 ng EE2/L. In the study, besides the three treatment groups, a control group is also employed, and we use L0, L1, L2 and L3 to represent the groups from control to highest level of treatment respectively. In addition, except L3 group animals whose gender is not clear, the other three groups have both male and female samples. A major outcome of the study is that the objects under highest level treatment tend to be female like. Therefore, an interesting question is that which genes expressed differently in common between the L3 subjects and other samples. Let L3/L0, L3/L1 and L3/L2 denote the expression ratio on log2 scale, then we want to know which genes are with significantly non zero values cross the three parameters. Due to p-values (from two-side test) for three parameters of each gene are available, both Fisher's method and our method are used to find such genes. Without loss generality, we simply take significance level $\alpha = 0.01$. The outcomes are that our method identified 2049 genes and Fisher's method identified 1831 genes. In the union of the two lists of genes, most of them are identified by both method, however, still a considerable number of them are

identified by only Fisher's method or our method. The results also shows that almost all the genes identified only by our test are genes whose three parameters are not close to zero. In contrast, if a gene is only identified by Fisher's methods, at least one of its three parameters has a value closing to zero. This is evident that our method performs better than Fisher's.

4 Conclusion

The comparative study of our method with Fisher's combined probability test confirm our method has a plausible property and stronger power. The point is shows by not only theoretical analysis, but also simulation results and application to practical cases.

The computation required by the proposed method becomes more and more intensive as the the increase of the number of p-values k. For very large k the application of our method is not as easy as Fisher's. Another problem is the availability of probability distribution function when $k > 3$, we hope there is a better way to get it in the future.

Funding

Y.F. was supported by NERC grants (NE/D000602/1) and NE/F001355/1.

Acknowledgement

We thank C. Tyler and L. Anke and A. Cossins, for providing and allowing to use their roach microarray data.

References

1. Fisher, R.A.: Statistical Methods for Research Workers, 4th edn. Oliver and Boyd (1932)
2. Tippett, L.H.C.: The method of Statistics. Williams and Norgate (1931)
3. Edgington, E.S.: An additive method for combining probability values from independent experiments. Journal of Psychology 80, 351–363 (1972)
4. Edgington, E.S.: A normal curve method for combining probability values from independent experiments. Journal of Psychology 82, 85–89 (1972)
5. Wilkinson, B.: A statistical consideration in psychological research. Psychological Bulletin 48, 156–157 (1951)
6. Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A., Williams Jr., R.M.: The American Soldier. In: Adjustment during Army Life, vol. I. Princeton University Press, Princeton (1949)
7. Birnbaum, A.: Combining independent tests of significance. Journal of the American Statistical Association 49, 559–574 (1954)
8. Littell, R.C., Folks, J.L.: Asymptotic optimality of Fisher's method of combining independent tests. Journal of the American Statistical Association 66(336), 802–806 (1971)

9. Littell, R.C., Folks, J.L.: Combining Independent Tests of Significance. Journal of the American Statistical Association 68(341), 193–194 (1973)
10. Berk, R.H., Cohen, R.: Asymptotically optimal methods of combining tests. Journal of the American Statistical Association 74, 812–814 (1979)
11. William, R.: A Consensus Combined P-Value Test and the Family-Wide Significance of Component Tests. Biometrics 46(2), 303–308 (1990)